DOCUMENT RESUME

ED 292 833 TM 011 211

AUTHOR Hackett, Rachelle Kisst; And Others
TITLE Test Construction Manipulating Score Differences between Black and White Examinees: Properties of the Resulting Tests. Research Report No. 87-30.
INSTITUTION Educational Testing Service, Princeton, NJ. Program Statistics Research Project.
PUB DATE Jul 87
NOTE 70p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Black Students; *Difficulty Level; *Item Analysis; Racial Bias; *Racial Discrimination; *Test Bias; Test Construction; Test Items; Test Reliability; White Students

ABSTRACT

This study was designed to investigate the empirical consequences of building tests for which the size of the difference in average scores between Black and White examinees is made to be as large and as small as possible within certain constraints. Test forms for two item types were developed that maximized or minimized the differences in average scores between Black and White examinees. The test forms were then administered to a random sample of examinees along with conventionally constructed control forms. The results suggest that impact manipulation is possible while maintaining content specifications and average difficulty level. However, the distribution of item difficulties was altered as were the reliability of the experimental sections and the relationships of these sections to other parts of the test. (Author/TJH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

RR-87-30

RESEARCH REPORT

# TEST CONSTRUCTION MANIPULATING SCORE DIFFERENCES BETWEEN BLACK AND WHITE EXAMINEES: PROPERTIES OF THE RESULTING TESTS

Rachelle Kisst Hackett
Paul W. Holland
Mari Pearlman
Dorothy T. Thayer

ED292833

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Helen C. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 011 211

Educational Testing Service
Princeton, New Jersey
July 1987

BEST COPY AVAILABLE

# TEST CONSTRUCTION MANIPULATING SCORE DIFFERENCES BETWEEN BLACK AND WHITE EXAMINEES: PROPERTIES OF THE RESULTING TESTS

Rachelle Kisst Hackett
Paul W. Holland
Mari Pearlman
Dorothy T. Thayer

Educational Testing Service

February, 1987

Copyright © 1987. Educational Testing Service. All rights reserved.

**ABSTRACT**

Test forms for two item types were developed that maximized or minimized the differences in average scores between Black and White examinees. The test forms were then administered to random samples of examinees along with conventionally constructed control forms. The results suggest that impact manipulation is possible while maintaining content specifications and average difficulty level. However, the distribution of item difficulties was altered as were the reliability of the experimental sections and the relationships of these sections to other parts of the test.

## PREFACE

This report and the research it describes is the result of the work of several people. The study was jointly designed by P. Holland, M. McPeek, M. Pearlman, and D. Thayer. The test development was supervised by M. Pearlman and carried out by P. Cooper and D. Eglewski. Data analyses were designed and carried out by P. Holland and D. Thayer. R. Hackett, an ETS summer intern, wrote the report in conjunction with P. Holland. Funding for this research was supplied by the ETS Program Research Planning Council as part of a larger study of differential item functioning directed by M. McPeek and C. Wild. The order of authorship for this report is alphabetical.

## Table of Contents


| | | Page |
|---|---|---|
| 1. | Introduction | 1 |
| 2. | Background | 3 |
| 3. | The Experiment | 5 |
| | 3.1 The Development of the Experimental Forms | 5 |
| | 3.2 The Administration of the Experimental and Control Sections | 9 |
| 4. | Black/White Score Differences on the Experimental and Control Sections | 10 |
| 5. | Test Development Review of the Experimental Sections | 12 |
| 6. | Statistical Characteristics of the Experimental and Control Sections | 14 |
| | 6.1 Difficulty Measures | 14 |
| | 6.2 Reliability Measures | 16 |
| | 6.3 Correlations with Other Variables | 18 |
| 7. | Three Other Analyses | 19 |
| | 7.1 Differential Item Functioning | 20 |
| | 7.2 Differential Omitting | 21 |
| | 7.3 Evidence for Multidimensionality in the Experimental and Control Sections | 24 |
| 8. | Conclusions | 26 |
| | References | 34 |
| | Appendix A: The Influence of Variation in Item Characteristics on the Variance of Test Scores | |

## 1. INTRODUCTION

This study was designed to investigate the empirical consequences of building tests for which the size of the difference in average scores between Black and White examinees is made to be as large and as small as possible within certain constraints. Our aim was to create tests that would adhere to operational content and statistical specifications but at the same time exhibit Black/White differences in scores as small and as large as possible under the constraints imposed by the test specifications and the size of our available item pool. To our knowledge this has never been done before, although related research may be found in: Flaugher and Schrader (1978), Green (1972), Ironson and Craig (1982), Kok et al. (1985) and Subkoviak et al. (1984).

Our motivation for this study stems from two sources. First of all, it is difficult to see how the study of so-called "biased" tests and test questions can proceed without examples of real tests (as opposed to artificial or make-believe tests) that are clearly biased. The terms "item bias" and "test bias" do not refer to precise concepts but rather to loosely connected collections of concerns about test fairness to all examinees. Thus "bias" is too vague a term to lend itself to direct experimental manipulation and study. A major concern involving test bias is the large score difference between (for example) Black and White examinees. These score differences <u>can</u> be experimentally manipulated, as our study shows, and lead to a clear, but not necessarily the only, definition of a biased test.

The tests produced in this study are biased in a very explicit way, although they do not appear so on their surface. Our "maximum impact tests" are <u>constructed</u> to have large Black/White differences in scores and are therefore "biased" against Black examinees in a clearly defined way. The "minimum impact

tests" have small Black/White differences in scores and may be interpreted in at least two ways. First, as tests that have been altered to be less biased against Black examinees and second, as tests that have been altered to be biased against White examinees. The first interpretation is based on the assumption that if the test were not biased then there would be no Black/White score difference. The second interpretation is based on the assumption that the observed Black/White difference in scores reflects legitimate differences in performance due to the quality of education and other relevant factors that are also distributed differently between these two groups of examinees. Having real examples of differentially biased tests and relevant data to evaluate them may help the study of test bias reach more definitive conclusions.

A second motive for this study stems from several recent legislative actions that are aimed at lessening Black/White (and other) differences in test scores by mandating test construction that selects test items, based on pretest data, to minimize these differences in accordance with the so-called "Golden Rule" procedure. In the legislative arena much has been argued both for and against this type of test construction (McAllister, 1987); however, there are few studies to inform us about the properties of tests constructed in this way. Do such tests change completely? Do the Black/White differences in scores become negligible? Do the tests become more difficult? Less difficult? More reliable? What are the costs involved? This study was designed to give some answers to these questions for two specific item types on a graduate-level admissions test (herein called the TEST). The information generated by studies such as the one reported here can serve to inform the debate over the wisdom of legislation that mandates test construction procedures aimed at minimizing group differences by statistical methods.

## 2. BACKGROUND

In this study comparisons are made between a focal group and a reference group of examinees. The focal group consists of Black examinees; White examinees serve as the reference group. The designations "Black and White" are self-reported by the examinees by marking the registration form. The overall rate of nonresponse of U.S. citizens to this question is about five percent. Only data from United States citizens are used in this study.

The "impact" of an item (test question) is the difference between the reference and the focal group in the proportion correctly answering the item, i.e.,

$$\text{IMPACT} = p_R - p_F.$$

(In this report we also use the synonymous term "p+ difference" to mean impact). In this study impact was manipulated in the construction of four experimental sections of the TEST, two Problem Solving (PS) sections and two Sentence Correction (SC) sections.

A Problem Solving section consists of 20 multiple choice mathematics questions whose solution requires facility with elementary arithmetic, algebra and geometry. A Sentence Correction section consists of 25 multiple choice questions in which a sentence is given that contains an underlined portion. The task is to replace the underlined portion with one of five alternative responses that "express most effectively what is presented in the original sentence." The first alternative response is always to leave the underlined portion of the sentence unchanged.

For each of the two item types, PS or SC, one experimental section was constructed to contain as many minimum impact items as possible from the item pool while the other experimental section was constructed to contain as many

maximum impact items as possible within test specification constraints. These four experimental sections will be designated herein as PSMin, SCMin, PSMax, and SCMax, respectively. Care was taken in the construction of these four sections to meet the usual content and statistical specifications of PS and SC sections as they are regularly constructed by ETS test developers. The two test developers who assembled the experimental sections each had several years of experience in the assembly of these types of sections.

In the March 1986 administration of the TEST, these four experimental sections were included in the nonoperational variable sections of the TEST. This yielded two types of data: (a) examinee responses to the four experimental sections, (b) examinee responses to various PS and SC sections that were assembled according to the normal procedures for such test sections. These latter "normal" sections serve as control sections for the study. There are two types of control sections, the operational test sections (two PS and one SC section) and four PS and two SC preoperational sections that were also included in the nonoperational variable sections of the March 1986 TEST for equating purposes. Due to the way that data are collected in the nonoperational sections of the TEST (i.e., spiraling) the examinees taking any of the four experimental sections or the six control preoperational sections may be regarded as random samples from the entire March 1986 TEST administration. This sampling design is important since it tends to equalize the distribution of ability of the examinees in the samples taking each of these sections. Because the reported scores of examinees' are based solely on their performance on the operational parts of the test, the results of their performance on the experimental sections did not count towards their test scores. Thus, the examinees' performance on minimum and maximum impact sections did not affect their reported scores.

As an additional check on the experimental sections, test developers not involved in this experiment were asked to examine three sections (either all PS or all SC sections) that included the maximum impact, the minimum impact, and a control section. Their instructions were to see if there were any questions in those sections that might advantage or disadvantage Black examinees. They were not told that the three sections they examined were different in any way.

The remainder of this report is organized as follows. Section 3 describes the experiment in detail. Section 4 compares the experimental and control forms in terms of score differences between the reference and focal group. Section 5 describes the results of the test development review of the experimental forms. In section 6 the experimental and control forms are compared in difficulty and reliability, their distributions of item characteristics and their correlations with other relevant variables. In section 7 the experimental and control sections are compared for evidence of differential item functioning, differential omitting, and their comparative dimensionality as assessed by item response theory. Our conclusions are then discussed in section 8.

## 3. THE EXPERIMENT

### 3.1 The Development of the Experimental Forms

Description of the Item Pool. During the summer of 1985 three tests which had been given preoperationally for pre-equating purposes but which had not been used operationally were identified. These are designated F1, F2, and F3 in this report. Each of the tests contains two Problem Solving sections (PS1 and PS2) composed of 20 items each and one Sentence Correction (SC) section composed of 25 items. Thus, pools of $3 \times 2 \times 20 = 120$ Problem Solving items and $3 \times 25 = 75$ Sentence Correction items were available to construct the experimental test sections.

An overview of the origin of the items used to develop the experimental sections is given below.

-- Pretest of items; pretest item statistics collected: delta and biserial correlation.

-- Assembly of the three forms F1, F2, and F3 based on content specifications and pretest statistics.

-- Preoperational administration of the forms F1, F2, and F3 for equating purposes, twice.
F1: 3/84, 10/84
F2: 10/84, 3/85
F3: 1/85, 6/85.

-- Assembly of experimental forms from the PS and SC item pools formed by combining forms F1, F2, and F3.

In August 1985 the Black/White impact for each item in the PS and SC sections of F1, F2, and F3 was analyzed. Impact was computed for each item and then averaged across its two preoperational administrations. The number of Black examinees responding to each item ranged from 250 to 350 across the three preoperational forms while the corresponding number of White examinees ranged from 3600 to 4700. Instead of using the actual p+ differences to measure impact, we grouped it into six categories labeled A, B, C, D, E and F. In Table 1 below, we give the range of p+ differences used in each impact category -- A, B, C, D, E, and F -- and give the distribution of items in these six categories.

---
Table 1 about here
---

The classification of items into the categories A through F was based on intervals developed in a pilot study. In the pilot study, using different items, there were approximately an equal number of items, across the two item types, in each category. This is approximately satisfied in the data of this study for categories A, B, and D, but category C contains more than twice as many items than does either E or F.

**TABLE 1. IMPACT CLASSIFICATION SCHEME AND DISTRIBUTION OF ITEMS IN THE TWO ITEM POOLS**

| Code | p+ difference* | PS Items | | SC Items | | Total |
|---|---|---|---|---|---|---|
| | | N | % | N | % | |
| A | ≤ .08 | 9 | 7.5 | 21 | 28.0 | 30 |
| B | .09 - .13 | 18 | 15.0 | 16 | 21.3 | 34 |
| C | .14 - .18 | 31 | 25.8 | 25 | 33.3 | 56 |
| D | .19 - .23 | 25 | 20.8 | 6 | 8.0 | 31 |
| E | .24 - .28 | 15 | 12.5 | 6 | 8.0 | 21 |
| F | ≥ .29 | 22 | 18.3 | 1 | 1.3 | 23 |
| | | 120 | 99.9 | 75 | 99.9 | 195 |

*Based on preoperational data.

The two item types differ markedly in their measured impact values, with SC items having noticeably fewer high impact items than the PS items. This difference appears to be characteristic of these two item types. Indeed over 82 percent of the SC item pool was in categories A, B and C, compared to 48 percent of the PS items. Thus, at the outset we see that the amount of impact manipulation that is possible depends on the item type and on the item pools.

The Experimental Forms. Test assemblers, experienced in the construction of PS and SC sections, were given the pools of items and instructed to create two test forms from each of them, a minimum and a maximum impact experimental form. They were further instructed to observe all of the usual constraints they observed when assembling PS or SC forms -- content coverage and statistical specifications. To construct the minimum impact form they were to use the items in the A, B, and C categories if they could and to only use the D, E, or F items if they could not meet the test specifications with the other items. The instructions were reversed for the assembly of the maximum impact sections. The distribution of the impact categories (A through F) of the items chosen for the four experimental sections are shown in Table 2.

---

Table 2 about here

---

Both minimum impact sections used only items from categories A, B, and C. Because of the distribution of the items in the pools over the impact categories the maximum impact section did require the use of some C and B items. Overall, however, there is a clear difference in the required direction between the minimum and maximum impact sections.

When interviewed after the assembly of the experimental sections both test developers indicated that the content and statistical specifications for both of

TABLE 2. THE NUMBER OF ITEMS IN EACH IMPACT CATEGORY FOR THE ITEMS IN THE EXPERIMENTAL SECTIONS*

| Impact Category | PSMin | | SCMin | | PSMax | | SCMax | |
|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | N | % | N | % |
| A | 4 | 20.0 | 18 | 72.0 | | | | |
| B | 6 | 30.0 | 6 | 24.0 | | | 4 | 16.0 |
| C | 10 | 50.0 | 1 | 4.0 | 2 | 10.0 | 12 | 48.0 |
| D | | | | | 5 | 25.0 | 3 | 12.0 |
| E | | | | | 5 | 25.0 | 5 | 20.0 |
| F | | | | | 8 | 40.0 | 1 | 4.0 |
| | 20 | | 25 | | 20 | | 25 | |

* Based on preoperational data.

the experimental sections they assembled were met with relative ease. However, they both thought that a third section would have been difficult to put together. For SC, this is not surprising since there were only enough SC items for three sections while for PS there were enough items for six sections. Both assemblers felt that the experimental sections were parallel to all other operational PS or SC sections they had assembled in the past two years.

To show how the experimental sections compared to the other PS and SC sections in terms of meeting the statistical specifications for the test, Table 3 gives the desired and achieved delta and biserial values for the 13 experimental and control sections used in this study. (Forms F4, F5 and F6 are discussed in section 3.2).

The ETS item difficulty index, the item delta, has a range, in practice, from about 6.0 (an easy item) to about 20.0 (a hard item) with 13.0 indicating middle difficulty.

---

Table 3 about here

---

Although the average delta values of the items used in the experimental sections are typical of those in other corresponding sections, the average biserial values just meet the statistical guidelines. The maximum impact sections have average biserials that fall at the high end of the desired interval while the average biserials of the minimum impact sections fall at the lower end of the interval. This is not a chance happening but rather it is a consequence of the relationship between the three item statistics: delta, biserial, and impact. Figures 1-4 show the scatter-plots of impact (vertical axis) versus delta (Figures 1 and 2) and biserials (Figures 3 and 4) computed on the preoperational data for the PS and SC item pools. These are simple deltas and

TABLE 3. STATISTICAL GUIDELINES AND ACHIEVED VALUES FOR EXPERIMENTAL AND CONTROL SECTIONS.

| | ACHIEVED* | | | ACCEPTABLE RANGE | | |
|---|---|---|---|---|---|---|
| | Delta Mean | S.D. | Biserial Mean | Delta Mean | S.D. | Biserial Mean |
| PS Sections | | | | | | |
| Experimental | | | | | | |
| PSMin | 11.5 | 2.1 | 0.46 | (11.4,11.5) | 2.1+ | (.46, .52) |
| PSMax | 11.5 | 2.2 | 0.52 | . | | . |
| Controls | | | | | | |
| F4 PS1 | 11.5 | 2.2 | 0.51 | . | | . |
| PS2 | 11.6 | 2.2 | 0.52 | . | | . |
| F5 PS1 | 11.6 | 2.2 | 0.51 | . | | . |
| PS2 | 11.4 | 2.2 | 0.53 | . | | . |
| F6 PS1 | 11.4 | 2.1 | 0.53 | . | | . |
| PS2 | 11.6 | 2.1 | 0.53 | . | | . |
| Control Average | 11.6 | 2.2 | 0.52 | | | |
| SC Sections | | | | | | |
| Experimental | | | | | | |
| SCMin | 11.8 | 2.6 | 0.44 | (11.7,12.0) | 2.0+ | (.44, .47) |
| SCMax | 11.7 | 2.0 | 0.48 | | | . |
| Controls | | | | . | | |
| F4 SC | 11.8 | 2.0 | 0.47 | . | | . |
| F5 SC | 11.8 | 2.0 | 0.46 | . | | . |
| F6 SC | 11.7 | 2.0 | 0.46 | . | | . |
| Control Average | 11.8 | 2.0 | 0.46 | | | |

* Values used at time of original test assembly of F4, F5, and F6 based on pretest data — see section 3.2 for identification of F4, F5 and F6.

biserials unadjusted for "not reached" items, and the biserial criterion score is the section score excluding the item. Each point in these graphs corresponds to an item. Although there is no linear trend between impact and delta there is a strong linear trend between impact and the biserial correlation. These scatterplots show why the average biserial coefficients for the items in the experimental sections fall at the extreme of the desired range. There is a clear conflict between choosing items with low impact and high biserials and between choosing items with high impact and low biserials. There is also a slight conflict between choosing high impact items and items with extremely high or low delta values. Except for SCMin these conflicts do not seem to have affected the assembly of the experimental sections as measured by pretest statistics. For SCMin, the standard deviation of the deltas is 2.6 compared to the control average of 2.0. This shows that the (pretest) deltas of SCMin are more spread out around their average than they are in the other SC sections. These trends will come up again in section 6.

---

Figures 1-4 about here

---

3.2 The Administration of the Experimental and Control Sections

The TEST consists of eight separately timed sections. The operational portion of the test, from which examinees' scores are obtained and reported, consists of two PS sections, one SC section, and three other sections. The remaining two sections are used for equating, pretesting new items, or for various experimental purposes, and do not count towards an examinee's reported score. The March 1986 administration included sections serving all three purposes. Twenty subforms were spiraled together for administration, resulting in a random sample, stratified by testing center, taking each of these 20 subforms.

FIGURE 1

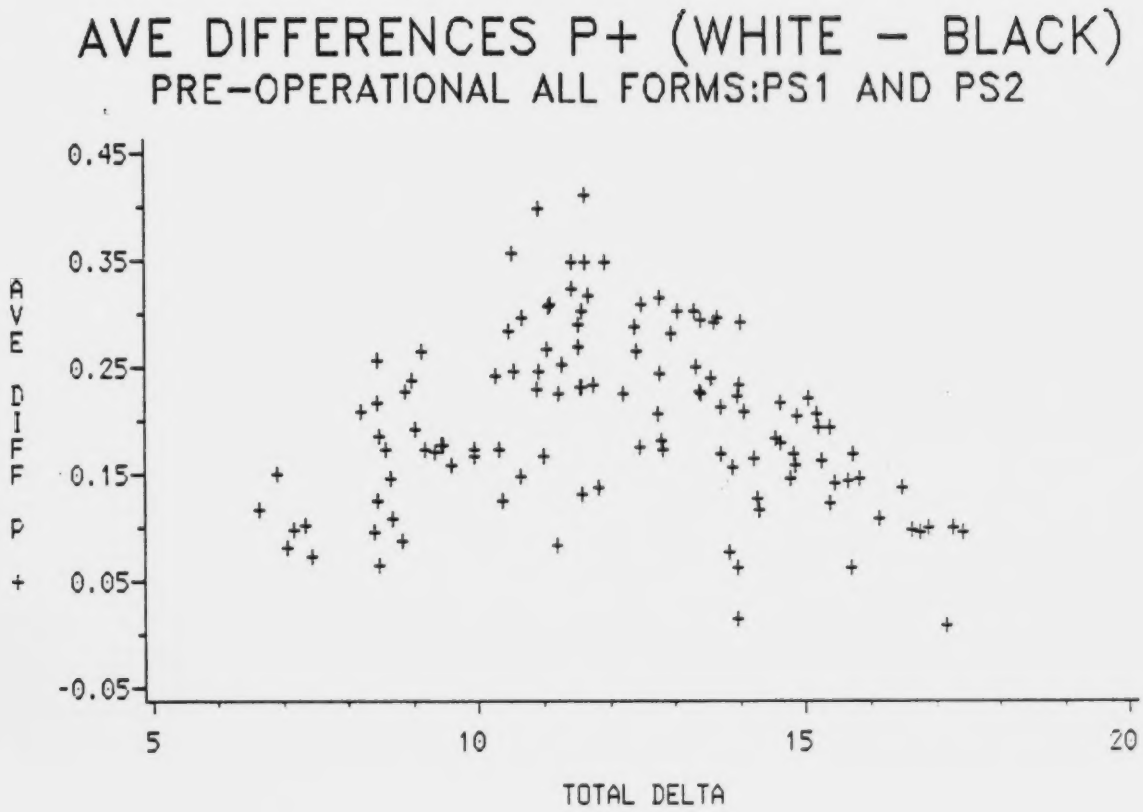

AVE DIFFERENCES P+ (WHITE - BLACK)
PRE-OPERATIONAL ALL FORMS:PS1 AND PS2
AVE DIFF P+
0.45
0.35
0.25
0.15
0.05
-0.05
5
10
15
20
TOTAL DELTA

**FIGURE 2**

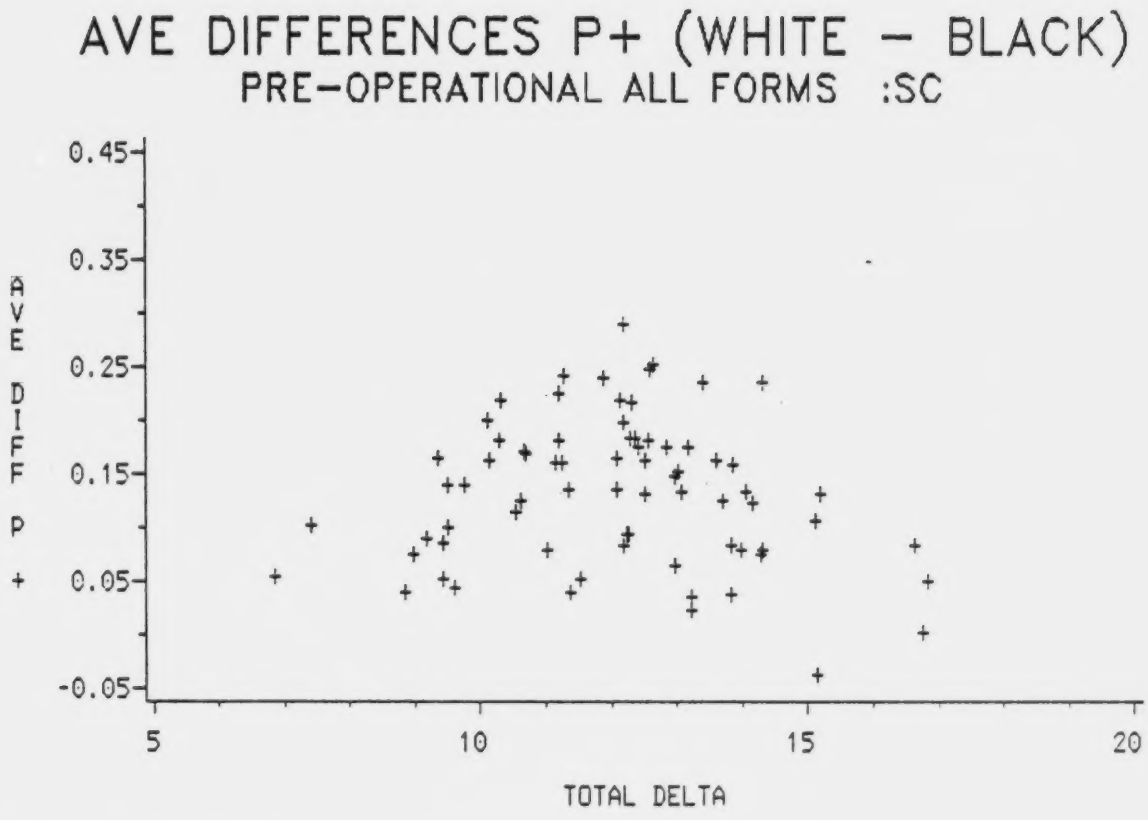

AVE DIFFERENCES P+ (WHITE – BLACK)
PRE-OPERATIONAL ALL FORMS :SC
AVE DIFF P+
0.45
0.35
0.25
0.15
0.05
-0.05
5
10
15
20
TOTAL DELTA

FIGURE 3

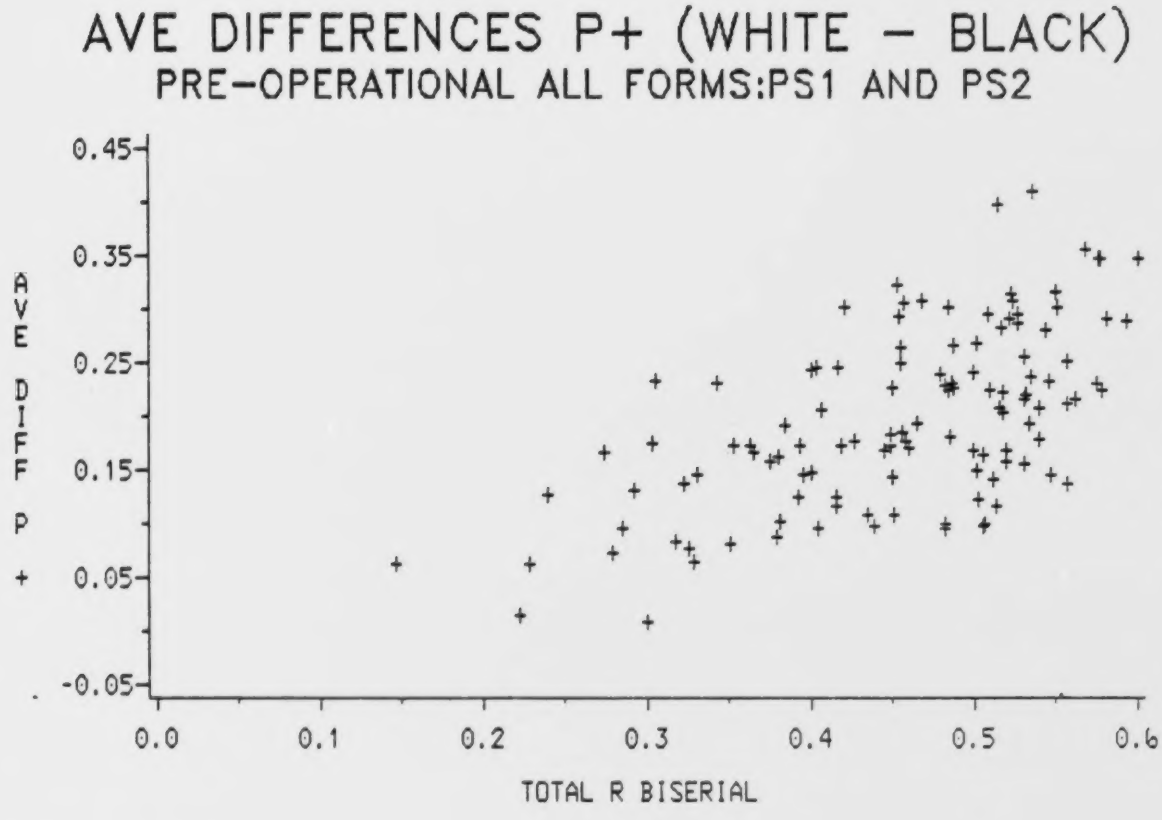
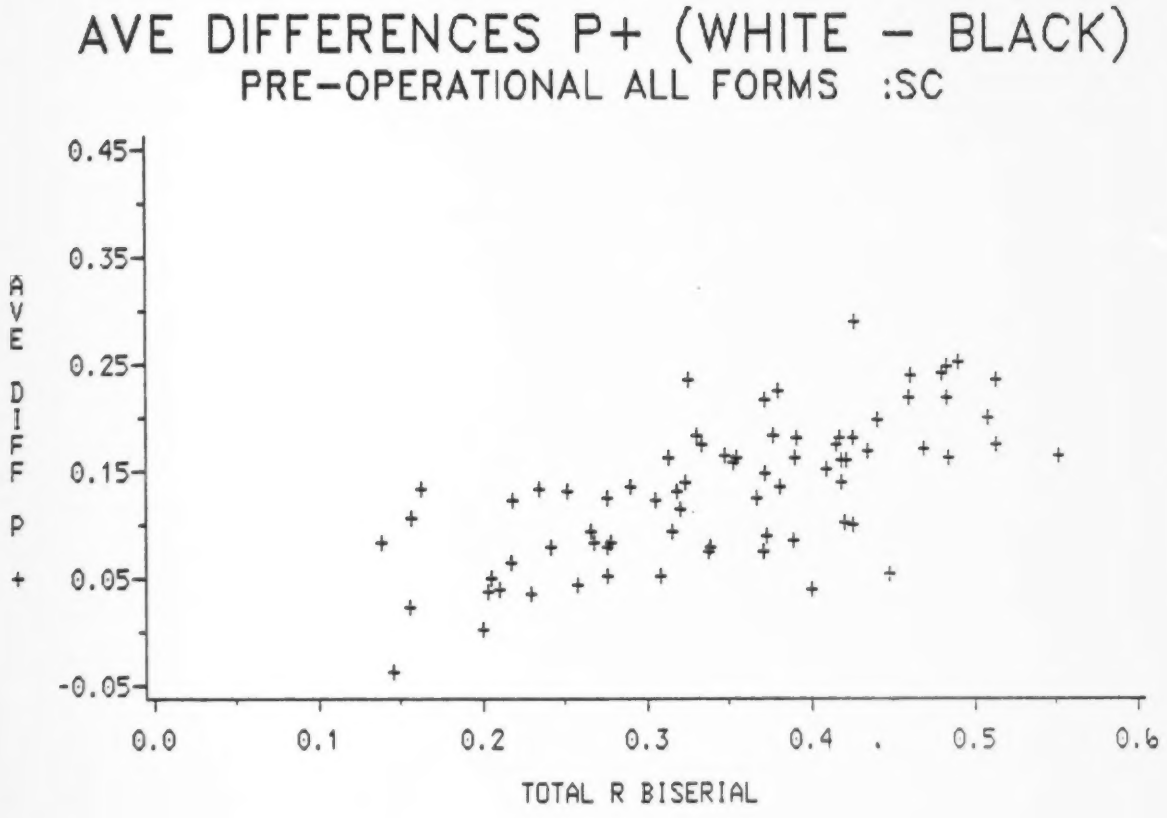

AVE DIFFERENCES P+ (WHITE – BLACK)
PRE-OPERATIONAL ALL FORMS:PS1 AND PS2
AVE DIF P+
0.45
0.35
0.25
0.15
0.05
-0.05
0.0
0.1
0.2
0.3
0.4
0.5
0.6
TOTAL R BISERIAL

FIGURE 4


AVE DIFFERENCES P+ (WHITE – BLACK)
PRE-OPERATIONAL ALL FORMS :SC
AVE DIFF P+
0.45
0.35
0.25
0.15
0.05
-0.05
0.0
0.1
0.2
0.3
0.4
0.5
0.6
TOTAL R BISERIAL

The two preoperational tests given, designated here by F5 and F6, provide four PS sections and two SC sections that we used as control sections along with the PS1, PS2, and SC sections of the operational test, designated here by F4. A summary of the sample sizes for the sections of interest in this study is provided in Table 4.

---

Table 4 about here

---

Although the number of Black examinees taking each experimental and preoperational form is not as large as recommended for studies of differential item functioning (which requires assessing impact conditionally given total score), these sample sizes are sufficient for us to evaluate the consequences of manipulating simple impact in test construction.

The control sections are all parallel in form and content to the corresponding experimental sections with the exception that the PS1 section of F4 contained only 19 items used in scoring, as opposed to 20 items in all of the other PS sections.

## 4. BLACK/WHITE SCORE DIFFERENCES ON THE EXPERIMENTAL AND CONTROL SECTIONS

Creating low and high impact test sections based on data from prior administrations of the items does not guarantee that they will exhibit low or high impact when given to a new sample of examinees unless the initial samples are large enough. However, our prior experience with this type of data and corresponding sample sizes gave us some assurance that the experimental sections would behave as predicted. To measure score differences between Black and White examinees on the experimental and control sections we used the following index:

$$\text{Standardized Score Difference} = \frac{\bar{X}_W - \bar{X}_B}{S_{W+B}} ,$$

TABLE 4. MARCH 1986 TEST SECTIONS ANALYZED IN THE STUDY AND CORRESPONDING SAMPLE SIZES.

| PS Sections | | Whites | Blacks | Others | Total |
|---|---|---|---|---|---|
| Experimental | | | | | |
| | PSMin | 1023 | 64 | 323 | 1410 |
| | PSMax | 1025 | 66 | 294 | 1385 |
| Controls | | | | | |
| Operational | F4 PS1 | 14833 | 1083 | 4522 | 20438 |
| | PS2 | 14833 | 1083 | 4522 | 20438 |
| Preoperational | F5 PS1 | 2167 | 153 | 636 | 2956 |
| | PS2 | 2184 | 136 | 638 | 2958 |
| | F6 PS1 | 2131 | 166 | 686 | 2983 |
| | PS2 | 2113 | 174 | 658 | 2945 |
| SC Sections | | | | | |
| Experimental | | | | | |
| | SCMin | 1023 | 64 | 323 | 1410 |
| | SCMax | 1025 | 66 | 294 | 1385 |
| Controls | | | | | |
| Operational | F4 SC | 14833 | 1083 | 4522 | 20438 |
| Preoperational | F5 SC | 2092 | 159 | 621 | 2872 |
| | F6 SC | 2095 | 168 | 620 | 2883 |

where $\bar{X}_W$, $\bar{X}_B$ denote averages for White and Black examinees and $S_{W+B}$ denotes the standard deviation for the combined sample of White and Black examinees. Table 5 gives values of this index for the experimental and control sections.

---

Table 5 about here

---

The results in Table 5 are clear enough. The experimental sections have standardized score differences in the predicted directions. Furthermore, there is no overlap of these values with the distribution of values for the control sections. Evidently the type of test construction used in this study can be used to manipulate the differences in score distributions for reference and focal group members. There are no substantial differences in these results between number right raw-scores or formula raw-scores. Since the TEST is formula scored, we shall use formula scores in the rest of this report (except Section 7). The PSMin standardized score difference is .22 less than the average for the control sections (a 20 percent decrease) and the PSMax standardized score difference is .30 more than the control average (a 28 percent increase). The SCMin standardized score difference is .27 less than the average for the control section (a 36 percent decrease) while the SCMax standardized score difference is .44 more than the control average (a 59 percent increase).

The standardized score differences are smaller for the SC sections mirroring the differences in Table 1. However, the effect of impact manipulation as a percent of the control average appears to be larger for the SC item type. Figure 5 graphically displays the data in Table 5.

---

Figure 5 about here

---

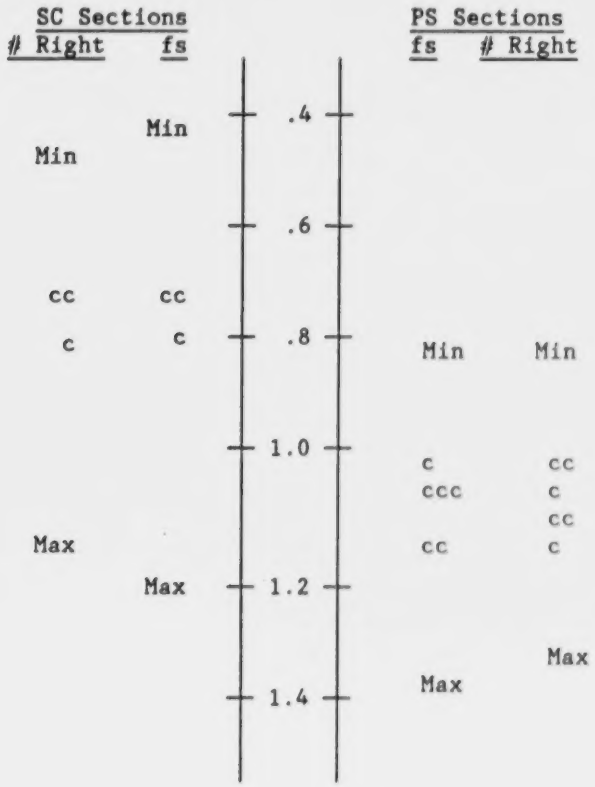
TABLE 5. STANDARDIZED SCORE DIFERENCES* FOR NUMBER RIGHT AND FORMULA SCORES FOR EXPERIMENTAL AND CONTROL SECTIONS.

| PS Sections | | Number Right | Formula Score |
|---|---|---|---|
| Experimental | | | |
| | PSMin | .88 | .87 |
| | PSMax | 1.31 | 1.39 |
| Controls | | | |
| | F4 PS1 | 1.04 | 1.06 |
| | PS2 | 1.01 | 1.04 |
| | F5 PS1 | 1.06 | 1.08 |
| | PS2 | 1.11 | 1.12 |
| | F6 PS1 | 1.10 | 1.14 |
| | PS2 | 1.08 | 1.08 |
| Control Average | | 1.07 | 1.09 |
| **SC Sections** | | | |
| Experimental | | | |
| | SCMin | .56 | .47 |
| | SCMax | 1.14 | 1.18 |
| Controls | | | |
| | F4 SC | .74 | .73 |
| | F5 SC | .81 | .79 |
| | F6 SC | .74 | .71 |
| Control Average | | .76 | .74 |

*Standardized Score Difference = $(\bar{X}_W - \bar{X}_B) \div S_{W+B}$

FIGURE 5

INSIDE OUT DISPLAY OF STANDARDIZED SCORE DIFFERENCES FOR NUMBER RIGHT (# Right) AND FORMULA SCORES (fs) FOR ALL SECTIONS




c = control section
Min = Minimum impact section
Max = Maximum impact section

Having seen that impact manipulation is possible and that the size of these effects is not trivial one might ask how the relative score increases (minimum impact sections) and score decreases (maximum impact sections) were achieved. Table 6 shows the means of the section formula scores for Black and White examinees. Figure 6 graphs this information. From Figure 6 and Table 6 we see that the PS and SC sections differ somewhat in the way in which the manipulation

---

Figure 6 and Table 6 about here

---

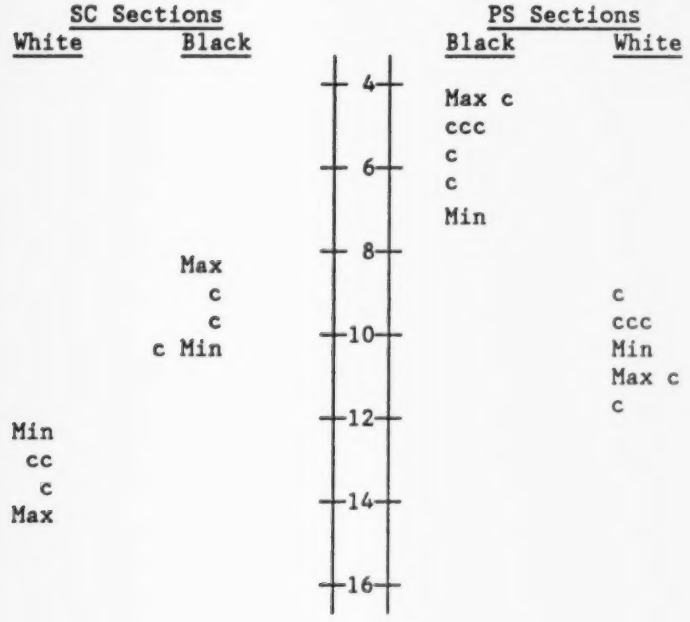
of impact was achieved. The mean formula scores on PSMin and PSMax for White examinees are near the control average for Whites of 10.13. Hence, for the PS sections the manipulation of impact was achieved primarily by altering the scores of Black examinees. However, for the SC sections impact manipulation was achieved by moving one group's scores up and the other group's scores down. The mean SCMin score for Blacks is greater than its control average while the mean SCMin score for Whites is less than its control average. The reverse is true for SCMax scores.

## 5. TEST DEVELOPMENT REVIEW OF THE EXPERIMENTAL SECTIONS

The instructions given to the test developers who assembled the experimental sections asked that they observe all of the constraints within which they normally assemble operational sections. Furthermore, they each reported that they regarded the two experimental sections as parallel to the corresponding operational test sections that they had assembled in the preceding two years. In order to independently verify that our experimental sections did not stand out as unusual upon an examination of their form and content we asked fourteen experienced ETS test developers to review the experimental sections and a control section according to the following design. Each participating test

FIGURE 6

INSIDE OUT DISPLAYS OF MEAN FORMULA SCORES FOR BLACK AND WHITE EXAMINEES FOR ALL SECTIONS




c = Control section
Max = Maximum impact section
Min = Minimum impact section

TABLE 6. MEAN OF SECTION FORMULA SCORES FOR BLACK AND WHITE EXAMINEES.

| | | Means | | S.D.(W+B) |
|---|---|---|---|---|
| PS Sections | | Blacks | Whites | |
| Experimental | | | | |
| | PSMin | 7.15 | 10.38 | 3.70 |
| | PSMax | 4.33 | 10.64 | 4.54 |
| Controls | | | | |
| | F4 PS1 | 5.46 | 9.80 | 4.07 |
| | PS2 | 5.48 | 9.71 | 4.07 |
| | F5 PS1 | 5.50 | 9.96 | 4.15 |
| | PS2 | 5.99 | 10.75 | 4.25 |
| | F6 PS1 | 6.25 | 11.31 | 4.43 |
| | PS2 | 4.31 | 9.22 | 4.54 |
| Control Average | | 5.50 | 10.13 | 4.25 |
| SC Sections | | | | |
| Experimental | | | | |
| | SCMin | 10.66 | 12.40 | 3.72 |
| | SCMax | 8.56 | 14.28 | 4.85 |
| Controls | | | | |
| | F4 SC | 9.62 | 12.69 | 4.20 |
| | F5 SC | 9.23 | 12.97 | 4.74 |
| | F6 SC | 10.43 | 13.52 | 4.38 |
| Control Average | | 9.76 | 13.06 | 4.44 |

reviewer was given three sections, all three either PS or SC depending on the expertise of the reviewer. These three sections consisted of the two experimental sections and a control section. The control section used for PS was PS1 from F5 while for SC we used F4. Seven test reviewers were given the three PS sections in different orders and seven were given the three SC sections, also in different orders. They were not told the ways in which these sections differed but were asked to review the items in each for their "potential differential impact on Black and White examinees." They were also asked to comment on the parallelism of the three forms.

The results of these reviews support the original test developers' views that in terms of their form and content the experimental forms are not different from the control forms. All of the reviewers said that the three forms they examined were reasonably parallel in content. Four of the seven PS reviewers could not find any basis for assessing any of the items on the three sections as having potential differential impact for Black and White examinees. The remaining three reviewers agreed on an assessment of differential impact on two items -- one in the control section and one in PSMax. We conclude that the reviewers saw little to differentiate the three PS sections.

For the SC sections, ten items were identified by three or more of the seven reviewers as having potential differential impact: six in the control section, three in SCMin and one in SCMax. Thus, if anything, the experimental SC sections appeared to these reviewers as having <u>fewer</u> potentially large Black/White impact items than did the control. Table 7 summarizes the results of these reviews.

---

Table 7 about here

---

TABLE 7. SUMMARY OF RESULTS OF THE TEST DEVELOPMENT REVIEWS** OF EXPERIMENTAL AND CONTROL SECTIONS.

| | PS Sections | | | SC Sections | | |
|---|---|---|---|---|---|---|
| | Control | PSMax | PSMin | Control | SCMax | SCMin |
| 1. Number of items flagged* by three or more reviewers | 1 | 1 | 0 | 6 | 1 | 3 |
| 2. Number of items flagged by one or more reviewers | 13 | 15 | 18 | 18 | 15 | 14 |
| 3. Total number of items flagged per section | 19 | 24 | 24 | 37 | 27 | 19 |
| 4. Number of SC items flagged due to possibly sensitive subject matter by two or more reviewers | -- | -- | -- | 8 | 3 | 6 |
| 5. Number of SC items flagged due to a grammatical construction believed by the reviewer to favor Whites by two or more reviewers | -- | -- | -- | 6 | 3 | 6 |

*flagged = identified by reviewers for "potentially differential impact on Black and White examinees."

**Seven reviewers each for the PS and the SC sections.

The reasons given by the three PS reviewers who were able to flag items did not fall into any clear pattern. However, the SC reviewers did use two common ways of flagging items. Some felt the subject matter of an item might appear to be more sensitive to Black than to White examinees. The other common reason given was the use of grammatical constructions that might favor White examinees. As we can see in Table 7 both of these potential problems were not found in SCMax as frequently as they were in the other two sections. Thus, while some agreement could be reached on the reasons for flagging SC items by the reviewers, the characteristics of flagged items do not appear to be associated with empirical impact.

## 6. STATISTICAL CHARACTERISTICS OF THE EXPERIMENTAL AND CONTROL SECTIONS

In this section we examine the statistical measures of difficulty and reliability for the experimental and control sections as well as their correlations with other measured variables.

### 6.1 Difficulty Measures

Table 8 and Figure 7 give the means and standard deviations of the item deltas for the experimental and control sections based on the March 1986 data.

---

Table 8 and Figure 7 about here

---

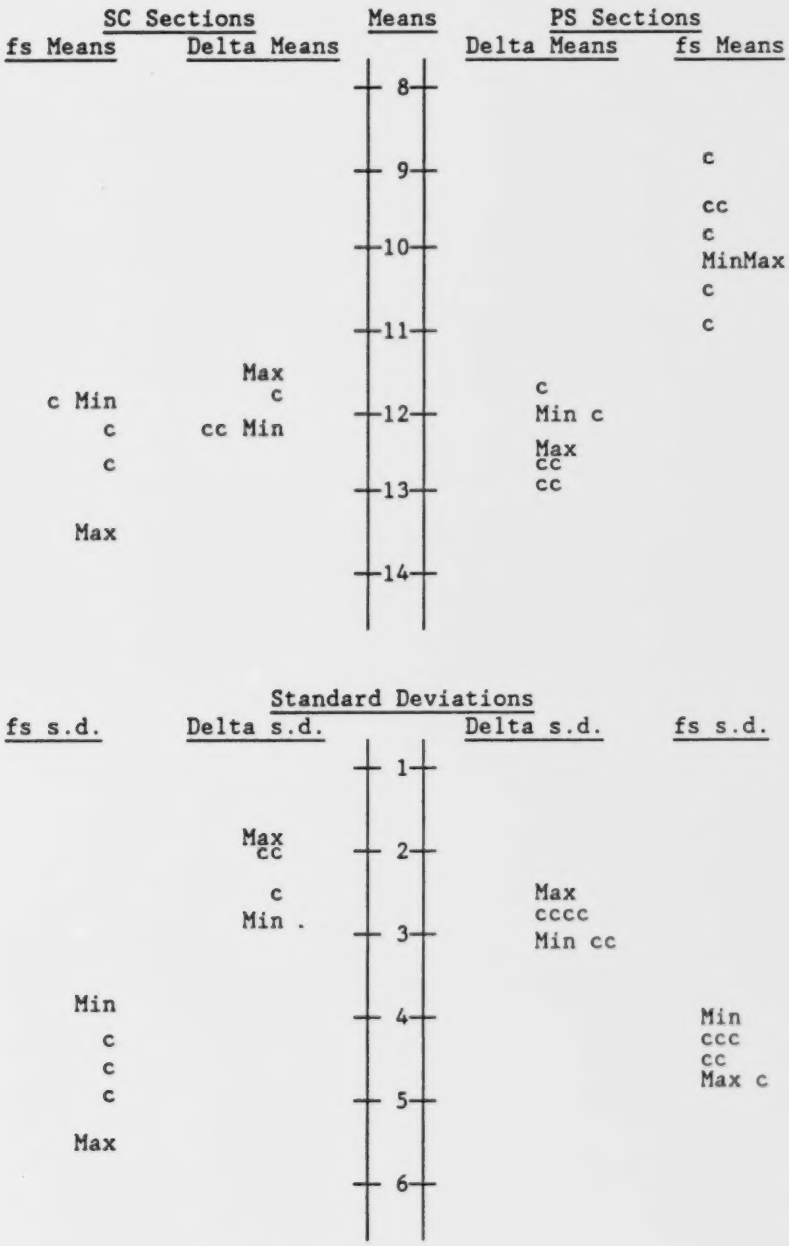
The mean deltas for the experimental sections are in the range of values given by the control sections except for SCMax whose mean delta of 11.7 is, nonetheless, only 0.4 less than the average of the control sections, 12.1. Thus, the mean deltas of the experimental sections are probably acceptably close enough for most purposes to the corresponding values for the control sections. However, the spread of the item difficulties for the experimental sections differs systematically from that of the controls. The maximum impact sections

TABLE 8. MEANS AND STANDARD DEVIATIONS OF ITEM DELTAS AND FORMULA SCORES FOR ALL GROUPS COMBINED.

| | | Item Deltas | | Formula Scores | | N |
|---|---|---|---|---|---|---|
| PS Sections | | Mean | s.d. | Mean | s.d. | |
| Experimental | | | | | | |
| | PSMin | 12.0 | 2.9 | 10.3 | 3.9 | 1410 |
| | PSMax | 12.3 | 2.5 | 10.2 | 4.6 | 1385 |
| Controls | | | | | | |
| | F4 PS1 | 12.4 | 2.7 | 9.6 | 4.2 | 20438 |
| | PS2 | 12.7 | 3.0 | 9.5 | 4.2 | 20438 |
| | F5 PS1 | 12.4 | 2.8 | 9.8 | 4.3 | 2956 |
| | PS2 | 12.0 | 2.7 | 10.5 | 4.4 | 2958 |
| | F6 PS1 | 11.8 | 2.6 | 11.0 | 4.5 | 2983 |
| | PS2 | 12.9 | 2.6 | 8.9 | 4.6 | 2945 |
| Control | Average | 12.4 | 2.7 | 9.9 | 4.4 | |
| SC Sections | | | | | | |
| Experimental | | | | | | |
| | SCMin | 12.1 | 2.9 | 11.9 | 4.0 | 1410 |
| | SCMax | 11.7 | 1.7 | 13.4 | 5.3 | 1385 |
| Controls | | | | | | |
| | F4 SC | 12.2 | 2.6 | 11.9 | 4.5 | 20438 |
| | F5 SC | 12.1 | 2.0 | 12.1 | 5.0 | 2872 |
| | F6 SC | 11.9 | 2.1 | 12.7 | 4.7 | 2883 |
| Control | Average | 12.1 | 2.2 | 12.3 | 4.8 | |

FIGURE 7

INSIDE OUT DISPLAYS OF MEANS AND STANDARD DEVIATIONS OF ITEM DELTAS AND SECTION FORMULA SCORES FOR ALL SECTIONS




c = Control section
Max = Maximum impact section
Min = Minimum impact section

have smaller variation in item deltas while the minimum impact sections have larger variation in item deltas than do typical control sections. This effect is stronger in SC than it is for PS, and is probably due to the relationship between impact and item deltas shown in Figures 1 and 2. There is a "triangular" trend in both of these graphs that causes high impact items to more typically have middle delta values than do low impact items. Thus, while the mean deltas are essentially the same for the experimental and control sections, the use of item impact to construct sections has changed the spread of item difficulty values in the experimental sections. High impact sections are therefore more "peaked" (i.e., have smaller variation in deltas) while low impact sections are less "peaked" in terms of their delta distributions. Except for SCMin, these trends were not observed in the experimental sections based on the pretest data that were available at their assembly -- see Table 3.

Table 8 also gives the means and standard deviation of the formula scores for all examinees on each experimental and control section. The mean formula scores reflect the results found for item difficulties -- the average scores of the experimental sections fall in or close to the range of values given by the control sections. The PS experimental sections fall into the control range while SCMax is somewhat easier than the control sections.

The standard deviations of the formula scores show a systematic trend in the comparison of experimental and control sections. The minimum impact sections have a smaller spread in their formula score distributions than do the control sections while the maximum impact sections have a larger spread. This relationship between impact manipulation and the standard deviations of formula score is what one might have predicted on the basis of the difference in the standard deviations of item deltas between the minimum and maximum impact sec-

tions. There is a general theoretical relationship between variation in difficulty across items and variation in test scores across examinees. Generally, a test with all equal deltas (a very peaked test) will spread out the scores of examinees more than will a test with variable deltas having the same average delta value (Gulliksen, 1945 and see Appendix A for a short exposition of this psychometric fact). In addition, these differences in formula score standard deviations are related to differences in section reliability, to which we now turn.

6.2 Reliability Measures

In section 3.1 we observed (Table 3) that while the average pretest biserial correlation coefficient for each experimental section is in or near the acceptable range, nonetheless there are clear trends for the minimum impact sections to have average biserial correlations at the low end of the range and for maximum impact sections to come in at the high end of the range. Table 3 is based on pretest data. Table 9 and Figure 8 are based on the data from the March 1986 administration, and Table 9 gives the mean and variance of the item biserial correlations as well as the section reliabilities, for all experimental and control sections.

---

Table 9 and Figure 8 about here

---

We see that the same trends that were seen in Table 3 also appear in Table 9. The mean biserials for the minimum impact sections fall below the range of values given by the control sections (i.e., .38 versus a control average of .46 for PSMin and .26 versus a control average of .33 for SCMin). The reverse is true for the maximum impact sections (i.e., .49 versus a control average of .46 for PSMax and .38 versus a control average of .33 for SCMax).

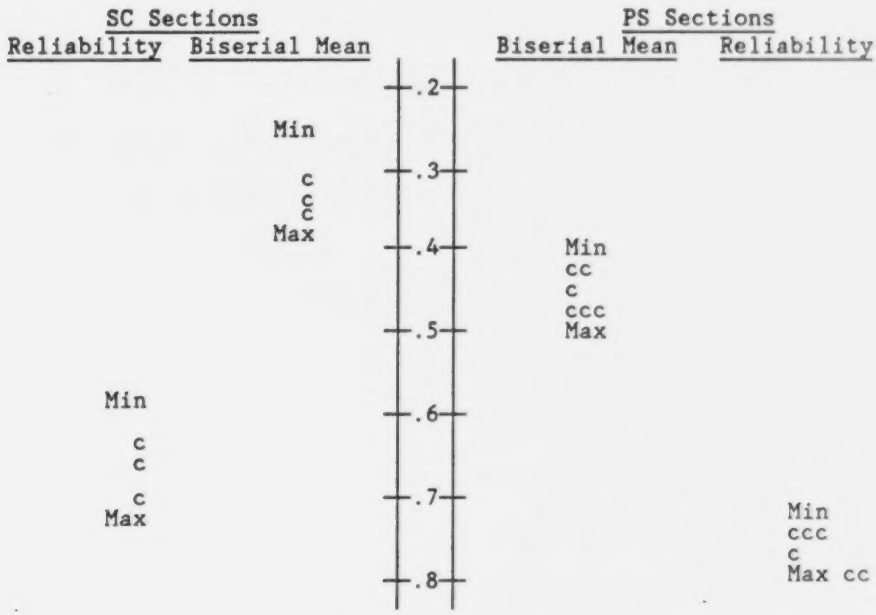
TABLE 9. MEAN AND STANDARD DEVIATION OF BISERIALS, AND RELIABILITY COEFFICIENTS FOR EXPERIMENTAL AND CONTROL SECTIONS

| | | Biserial Correlation* | | Reliability |
|---|---|---|---|---|
| PS Sections | | Mean | s.d. | (α-coefficient) |
| Experimental | | | | |
| | PSMin | .38 | .10 | .71 |
| | PSMax | .49 | .07 | .79 |
| Controls | | | | |
| | F4 PS1 | .43 | .12 | .76 |
| | PS2 | .44 | .07 | .75 |
| | F5 PS1 | .45 | .06 | .76 |
| | PS2 | .46 | .09 | .77 |
| | F6 PS1 | .49 | .08 | .79 |
| | PS2 | .47 | .07 | .79 |
| Control Average | | .46 | .08 | .77 |
| SC Sections | | | | |
| Experimental | | | | |
| | SCMin | .26 | .10 | .58 |
| | SCMax | .38 | .09 | .74 |
| Controls | | | | |
| | F4 SC | .32 | .09 | .67 |
| | F5 SC | .35 | .09 | .72 |
| | F6 SC | .33 | .07 | .68 |
| Control Average | | .33 | .08 | .69 |

*Criterion score = section total excluding item.

FIGURE 8

INSIDE OUT DISPLAY OF MEAN BISERIAL CORRELATION COEFFICIENTS AND RELIABILITY COEFFICIENTS FOR ALL SECTIONS




c = Control section
Max = Maximum impact section
Min = Minimum impact section

The differences in the average biserials for the Min and Max sections in Table 9 are larger than the corresponding differences at assembly (based on pre-test data) in Table 3. For PS the difference is .49 - .38 = .11 in Table 9 but only .52 - .46 = .06 in Table 3. For SC the difference is .38 - .26 = .12 for Table 9 but only .48 - .44 = .04 in Table 3. Part of these discrepancies is due to the inclusion of the item in the criterion in Table 3 and its exclusion from the criterion in Table 9. The relevant biserials in Table 9 were recomputed including the item and the difference becomes .07 for PS and .08 for SC. These are still larger than the difference in mean biserials that existed at assembly. This comparison is further confounded by other differences in the criteria used in Tables 3 and 9 to compute the biserials so it is not easy to decide whether tighter control on the average biserials at test assembly would have addressed the difference in the average biserials of the Min and Max sections seen in Table 9.

Table 9 also gives section reliabilities (coefficient $\alpha$) for the experimental and control sections. Because there is a close connection between biserial correlations and test reliability, it is not surprising that the section reliabilities show a pattern similar to the average biserials. The minimum impact sections have lower values of $\alpha$ than the control sections while the maximum impact sections have values of $\alpha$ that are at the top of or exceed the range given by the control sections.

One question that naturally arises is whether or not the reliability of the experimental sections shows the same pattern for Black and White examinees. Table 10 and Figure 9 give values of coefficient $\alpha$ and the standard deviation of formula scores separately for Black and White examinees. In considering Table 10 remember that the number of Black examinees is as low as 64 for PSMin and

SCMin -- see Table 4. The control sections are approximately equally reliable for Black and White examinees (i.e., control averages of .72 and .74 for PS sections and of .64 and .63 for SC sections). The minimum impact sections are less reliable for both Black and White examinees than are the control sections, and

---

Table 10 and Figure 9 about here

---

the maximum impact sections are slightly more reliable for both groups.

6.3 Correlations With Other Variables

There were two types of variables available to us with which we could correlate the scores on the experimental and control sections to see if the manipulation of impact would change the way that test scores are related to other relevant variables. These were: a) the other section scores on the TEST, (i.e., Data Sufficiency (DS), Analysis of Situations (AS), and Reading Comprehension (RC); and b) self-reported undergraduate grade point average (UGPA). Table 11 and Figure 10 summarize these correlations and the sample sizes for the correlations with UGPA.

The trends are very consistent in Table 11. The minimum impact sections are always less correlated with the operational section scores than are the control sections and the maximum impact sections are typically more highly correlated with the operational section scores than are the controls.

---

Table 11 and Figure 10 about here

---

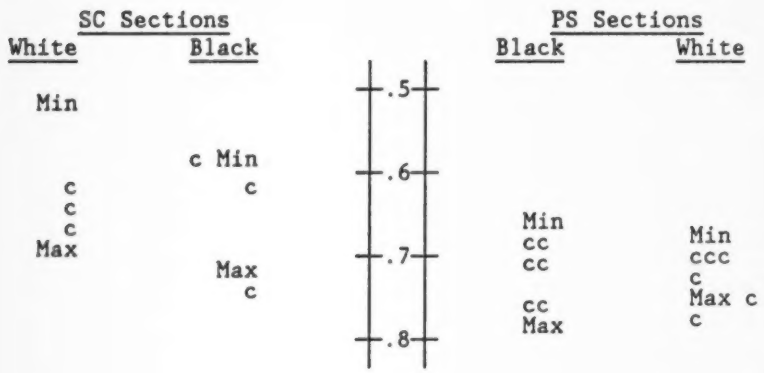
The correlations of the experimental and control sections with UGPA are substantially smaller than the other correlations in Table 11 -- ranging from .16 to .27. The UGPA-correlations for the experimental sections fall in the range specified by the control sections except for SCMax which, at .22, lies

TABLE 10. RELIABILITY COEFFICIENTS AND FORMULA SCORE STANDARD DEVIATIONS FOR BLACK AND WHITE EXAMINEES FOR EXPERIMENTAL AND CONTROL SECTIONS.

| | | Reliability (Coefficient α) | | Formula Score Standard Deviations | |
|---|---|---|---|---|---|
| PS Sections | | Black | White | Black | White |
| Experimental | | | | | |
| | PSMin | .65 | .66 | 3.6 | 3.6 |
| | PSMax | .78 | .76 | 4.6 | 4.3 |
| Controls | | | | | |
| | F4 PS1 | .72 | .73 | 4.0 | 3.9 |
| | PS2 | .67 | .73 | 3.6 | 4.0 |
| | F5 PS1 | .76 | .72 | 4.2 | 4.0 |
| | PS2 | .71 | .74 | 4.0 | 4.1 |
| | F6 PS1 | .76 | .77 | 4.4 | 4.2 |
| | PS2 | .70 | .76 | 3.7 | 4.4 |
| Control Average | | .72 | .74 | 4.0 | 4.1 |
| SC Sections | | | | | |
| Experimental | | | | | |
| | SCMin | .58 | .52 | 4.0 | 3.7 |
| | SCMax | .71 | .67 | 5.2 | 4.6 |
| Controls | | | | | |
| | F4 SC | .58 | .61 | 4.1 | 4.1 |
| | F5 SC | .72 | .66 | 5.1 | 4.6 |
| | F6 SC | .61 | .62 | 4.4 | 4.3 |
| Control Average | | .64 | .63 | 4.5 | 4.4 |

FIGURE 9

INSIDE OUT DISPLAY OF RELIABILITY COEFFICIENTS FOR BLACK AND WHITE EXAMINEES FOR ALL SECTIONS.



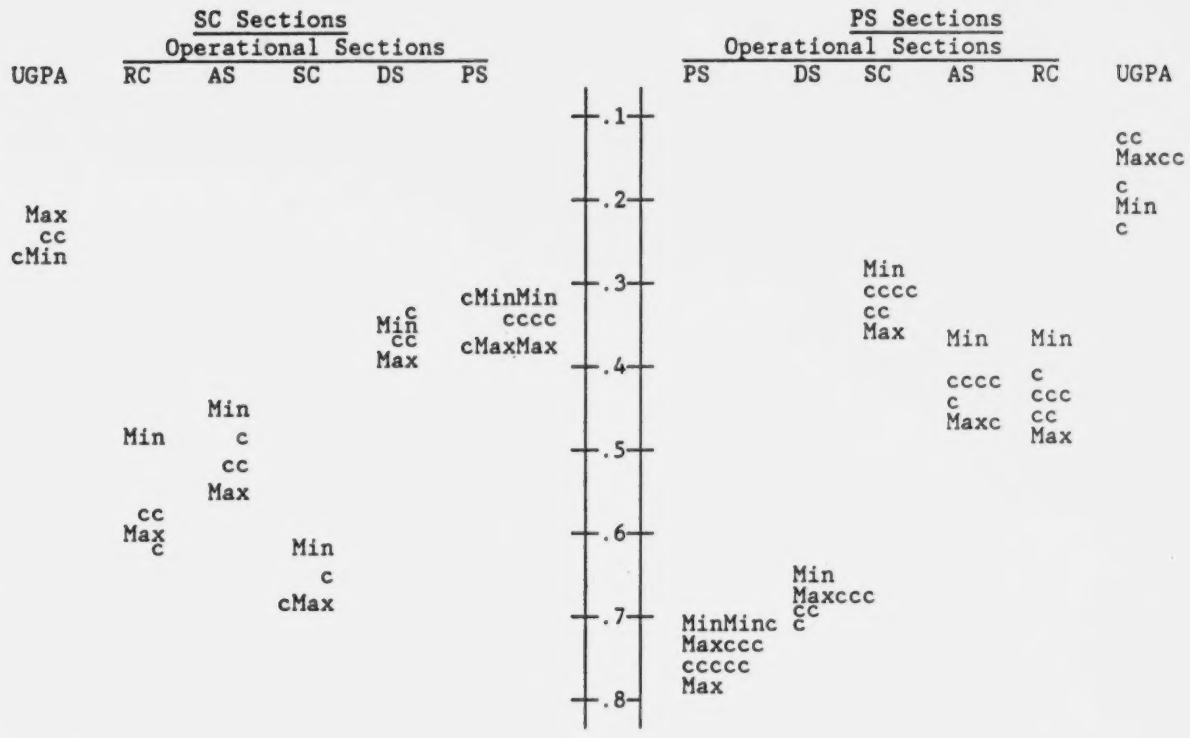

c = Control section
Max = Maximum impact section
Min = Minimum impact section

TABLE 11. CORRELATIONS OF FORMULA SCORES ON THE EXPERIMENTAL AND CONTROL SECTIONS WITH THE OPERATIONAL FORM (F4) SECTION SCORES AND UGPA.

| | | Operational Sections | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| PS Sections | | PS1 | PS2 | SC | DS | AS | RC | UGPA | N for UGPA |
| Experimental | | | | | | | | | |
| | PSMin | .72 | .71 | .28 | .65 | .36 | .36 | .21 | 1149 |
| | PSMax | .77 | .75 | .35 | .68 | .44 | .46 | .18 | 1109 |
| Controls | | | | | | | | | |
| | F4 PS1 | --- | .76 | .31 | .69 | .42 | .40 | .18 | 16474 |
| | PS2 | .76 | --- | .32 | .67 | .41 | .41 | .17 | 16474 |
| | F5 PS1 | .75 | .72 | .33 | .67 | .43 | .41 | .22 | 2371 |
| | PS2 | .76 | .75 | .34 | .67 | .42 | .43 | .18 | 2359 |
| | F6 PS1 | .77 | .76 | .31 | .69 | .42 | .43 | .19 | 2396 |
| | PS2 | .76 | .75 | .32 | .71 | .44 | .42 | .16 | 2365 |
| Control Average | | .76 | .75 | .32 | .68 | .42 | .42 | .18 | |
| SC Sections | | | | | | | | | |
| Experimental | | | | | | | | | |
| | SCMin | .31 | .31 | .63 | .38 | .48 | .50 | .27 | 1149 |
| | SCMax | .36 | .36 | .68 | .41 | .54 | .61 | .22 | 1109 |
| Controls | | | | | | | | | |
| | F4 SC | .31 | .32 | --- | .36 | .50 | .58 | .25 | 16474 |
| | F5 SC | .35 | .36 | .68 | .40 | .52 | .63 | .27 | 2302 |
| | F6 SC | .35 | .35 | .67 | .39 | .52 | .58 | .24 | 2367 |
| Control Average | | .34 | .34 | .68 | .38 | .51 | .60 | .25 | |

FIGURE 10

INSIDE OUT DISPLAY OF CORRELATIONS WITH OPERATIONAL SECTION SCORES AND UGPA FOR ALL EXPERIMENTAL AND CONTROL SECTIONS




c = Control section
Max = Maximum impact section
Min = Minimum impact section

slightly below the smallest UGPA-correlation for the control sections, .24. For both PS and SC the UGPA-correlations are slightly larger for the Min sections than for the Max sections. The differences are small, and in the case of PS well within the range of correlation differences exhibited by the control sections.

The UGPA values used in these analyses are self-reported and are subject to substantial nonresponse. Comparing the N's in Tables 4 and 11 we see that the nonresponse for UGPA is about 20% across all the forms.

It should be noted that these correlations with undergraduate grades should not be confused with predictive validity coefficients -- i.e., correlations with grades in appropriate graduate programs. Because of the great diversity of meaning attached to undergraduate grades (different colleges and different undergraduate majors) it is difficult to interpret correlations of test scores with UGPA across colleges and majors. Graduate-level predictive validity studies usually focus on correlations between test scores and grades within a graduate school and a department. Because of their ambiguous interpretability, the substantial nonresponse and their small size we did not pursue more detailed analyses of the UGPA-correlations.

## 7. THREE OTHER ANALYSES

This section describes the results of three analyses that bear on interesting questions that arose in our study. They all suffer from the same problem -- sample sizes that are too small to yield definitive results. Nevertheless, we include them here for completeness and as suggestive possibilities for future studies. The first is concerned with assessing evidence of differential item functioning as measured by the Mantel-Haenszel procedure (Holland, 1985; Holland and Thayer, 1986). The second also used the

Mantel-Haenszel procedure, but addressed the issue of differential omitting between White and Black examinees on the experimental sections. The third analysis uses the Mantel-Haenszel procedure to investigate the dimensionality of the experimental and control sections, as suggested by Rosenbaum (1984). Some authors have argued that test and item bias are only possible with multidimensional tests -- see, for example, Linn and Harnisch (1981).

7.1 Differential Item Functioning

Several procedures for the study of differential item functioning (DIF) have been proposed. We used the Mantel-Haenszel procedure as described in Holland and Thayer (1986) to control for the ability differences in the groups being compared. Briefly, in this procedure examinees are grouped into comparable subsets based on their operational score for the relevant operational section -- PS1+PS2 for PS sections and SC for SC sections. For a given item this results in a 2×2×K array of data of the form

**Item Score**

| | Correct | Incorrect | |
|---|---|---|---|
| Reference Group | $A_j$ | $B_j$ | $n_{Rj}$ |
| Focal Group | $C_j$ | $D_j$ | $n_{Fj}$ |
| | $m_{1j}$ | $m_{0j}$ | $T_j$ |

$j = 1, \ldots, K.$

Here K is the number of score intervals used to define comparable examinees. $A_j$ is the number in the reference group who were in the $j^{th}$ matched group on ability and who answered the item right. $B_j$, $C_j$, and $D_j$ have corresponding interpretations. In the Mantel-Haenszel procedure these data are combined into a single overall index, the MH Delta Difference, MH D-DIF, that measures the

degree more difficult that reference group members find the item compared to comparable focal group members. This degree of difficulty is expressed in the ETS delta scale. A MH D-DIF value of 0 indicates no difference between the reference and focal group, while a negative MH D-DIF value indicates that the focal group found the item harder than did comparable reference group members. We used the technique suggested in Holland and Thayer (1986) of adding the studied item to the matching criterion to eliminate the statistical bias that occurs in DIF statistics when the studied item is not a part of the matching criterion.

Table 12 gives the mean MH D-DIF values and other related statistics for the experimental and preoperational control sections. The individual item MH D-DIF values are very unstable with such small sample sizes for the focal group, so we examine the more reliable average values across all the items in the section.

---

Table 12 about here

---

It is clear that the minimum impact sections do exhibit DIF in the direction of favoring the focal group (positive values) while the maximum impact sections exhibit DIF in the direction of favoring the reference group.

7.2 Differential Omitting

The TEST is formula scored and examinees are not encouraged to randomly guess when they are unsure of the correct response to an item. Consequently, examinees omit* a fair number of responses on this test, though it varies by

---

* By "omits" we include any item that is not responded to regardless of whether it is an "embedded omit" or a "not reached" item. We did this because there were so few omits (especially in SC, see Table 13) that a more refined analysis did not seem warranted.

TABLE 12. SUMMARY STATISTICS FOR MH D-DIF* VALUES FOR DIFFERENTIAL ITEM FUNCTIONING IN EXPERIMENTAL AND PREOPERATIONAL CONTROL SECTIONS.

| | | Mean | S.D. | Min | Max | $n_R$ | $n_F$ |
|---|---|---|---|---|---|---|---|
| PS Sections | | | | | | | |
| Experimental | | | | | | | |
| | PSMin | .32 | 1.2 | -2.1 | 3.0 | 1029 | 64 |
| | PSMax | -.45 | 1.1 | -3.7 | 2.0 | 1027 | 66 |
| Controls | | | | | | | |
| | F5 PS1 | -.16 | 0.7 | -1.9 | 1.2 | 2178 | 153 |
| | PS2 | -.27 | 0.8 | -2.1 | 1.2 | 2191 | 136 |
| | F6 PS1 | -.11 | 0.6 | -1.7 | 0.8 | 2140 | 166 |
| | PS2 | -.11 | 0.6 | -1.4 | 0.9 | 2121 | 174 |
| Control Average | | -.16 | 0.7 | | | | |
| SC Sections | | | | | | | |
| Experimental | | | | | | | |
| | SCMin | .39 | 1.2 | -2.1 | 3.0 | 1029 | 64 |
| | SCMax | -.59 | 0.9 | -2.5 | 1.5 | 1027 | 66 |
| Controls | | | | | | | |
| | F5 SC | .04 | 0.6 | -1.5 | 0.7 | 2095 | 159 |
| | F6 SC | -.26 | 0.8 | -3.0 | 1.1 | 2106 | 168 |
| Control Average | | -.11 | 0.7 | | | | |

*Matching criterion is the relevant operational section number right score plus the studied items' score.

item type. For example, out of 40 PS items an average of 10.0 were omitted on the operational part of the TEST. For SC there were 1.6 omits out of 25 items on the same operational test.

Table 13 gives the mean number of omits for Black, White and all examinees for the experimental and control sections.

---

Table 13 about here

---

For the experimental PS sections both White and Black examinees omitted less than they typically did for the control sections. For the experimental SC sections there was a difference between White and Black examinees. White examinees omitted less on both of the experimental sections than on the controls. Black examinees omitted less on SCMax than on the control sections, but this is not true for SCMin. These results suggest that omitting may play a role in impact manipulation. The PS results, for which there is the most omitting, suggest that both impact manipulation and maximization are associated with less omitting than occurs in the control sections. One might have predicted less omitting by Black examinees on PSMin (compared to their omitting on control sections) and more omitting on PSMax, but this did not occur. For SCMin and SCMax the reverse holds. We conclude that the effects of impact manipulation on the omitting behavior of examinees is complex and worthy of deeper study.

Comparing the omitted responses of Black and White examinees is complicated by the fact that lower scoring examinees omit at different rates than do higher scoring examinees. For example, in Table 13 Black examinees omit at higher rates than do White examinees. To eliminate this effect we did a Mantel-Haenszel type of analysis in which the data for the comparable groups of

TABLE 13. AVERAGE NUMBER OF OMITTED RESPONSES PER SECTION FOR BLACK, WHITE, AND ALL EXAMINEES FOR EXPERIMENTAL AND CONTROL SECTIONS.

| PS Sections | | Black | White | Total* |
|---|---|---|---|---|
| Experimental | | | | |
| | PSMin | 5.0 | 3.6 | 3.6 |
| | PSMax | 5.4 | 4.1 | 4.2 |
| Controls | | | | |
| | F4 PS1 | 5.8 | 4.6 | 4.6 |
| | PS2 | 6.9 | 5.4 | 5.4 |
| | F5 PS1 | 6.1 | 4.3 | 4.4 |
| | PS2 | 5.8 | 4.1 | 4.2 |
| | F6 PS1 | 5.8 | 4.1 | 4.3 |
| | PS2 | 6.8 | 4.9 | 5.0 |
| Control Average | | 6.2 | 4.6 | 4.7 |
| SC Sections | | | | |
| Experimental | | | | |
| | SCMin | 2.6 | 0.9 | 1.3 |
| | SCMax | 1.1 | 1.0 | 1.3 |
| Controls | | | | |
| | F4 SC | 2.2 | 1.3 | 1.6 |
| | F5 SC | 2.4 | 1.3 | 1.6 |
| | F6 SC | 2.6 | 1.2 | 1.6 |
| Control Average | | 2.4 | 1.3 | 1.6 |

* Includes Black, White, and "Other" examinees.

examinees were arrayed as follows.

| | Omitted the Item? Yes | No | |
|---|---|---|---|
| Reference Group | $A_j$ | $B_j$ | $n_{Rj}$ |
| Focal Group | $C_j$ | $D_j$ | $n_{Fj}$ |
| | $m_{1j}$ | $m_{0j}$ | $T_j$ |

$j = 1, \ldots, K.$

Instead of adjusting the overall differences in correct answer rates for the differences in ability distributions of the Reference and Focal groups this use of the Mantel-Haenszel procedure allowed us to adjust the overall differences in omitting rates for these differences in the ability distributions. However, because of the small number of examinees in the focal group we did not pursue this analysis at the individual item level. Table 14 gives the mean MH D-DIF values across all items in a section for the experimental and preoperational control sections. It should be noted that with this use of the Mantel-Haenszel procedure, a negative MH D-DIF value means that the focal group omits the item at a lower rate than do comparable reference group members.

---

Table 14 about here

---

The results of Table 14 do not show a strong pattern, but they suggest that the focal group members were less likely to omit items in the maximum impact sections than in the minimum impact or control sections when compared to comparably scoring reference group members. In addition, Black examinees were more likely to omit items on SCMin than were comparable White examinees. We are quite hesitant to make much of these results due to the small sizes of the focal

TABLE 14. MH D-DIF VALUES FOR DIFFERENTIAL OMITTING FOR EXPERIMENTAL AND CONTROL SECTIONS.

| | | MH D-DIF* Mean | Number of Items with Statistically Significant (.01) MH D-DIF Values in Indicated Direction + | − |
|---|---|---|---|---|
| PS Sections | | | | |
| Experimental | | | | |
| | PSMin | -.07 | 1 | 1 |
| | PSMax | -.71 | 0 | 1 |
| Controls | | | | |
| | F5 PS1 | -.23 | 0 | 0 |
| | PS2 | .01 | 2 | 0 |
| | F6 PS1 | -.39 | 0 | 2 |
| | PS2 | .14 | 0 | 0 |
| SC Sections | | | | |
| Experimental | | | | |
| | SCMin | 1.80** | 2 | 0 |
| | SCMax | -.60** | 0 | 0 |
| Controls | | | | |
| | F5 SC | .12 | 3 | 0 |
| | F6 SC | 1.50 | 4 | 0 |

*Item scored 1 if it is omitted or not reached and 0 otherwise. Matching criterion is operational PS or SC number right score.

**These are averages of the finite MH D-DIF values only, for some items there were undefined MH D-DIF values due to the sample sizes and low omit rates.

groups and the overall rarity of omitting.

7.3 Evidence for Multidimensionality in the Experimental and Control Sections

One school of thought argues that in order for an item to be biased against a group of examinees, the item must be multidimensional in the sense of measuring more than one latent trait of examinees (Linn and Harnisch, 1981). When items are multidimensional in this sense, and if the joint distributions of these multiple latent traits are not the same in two subpopulations of examinees, then the ranking of examinees on scores based on these items may be misleading if only one of the traits the item measures is of primary interest. For example, suppose a mathematics test is full of "story problems" written in English, and the test is used to rank examinees on mathematical proficiency. It is clear that if examinees also vary in their ability to read English (say a mix of foreign graduate students) then this test is clearly biased against those students whose English is weak, regardless of their mathematical proficiency. In most real examples multidimensionality cannot be assessed by visual inspection of the items. For this reason we used a statistical technique.

We wanted to see if our experimental and control sections differed in how "unidimensional" they were. Because the March 1986 data are all from random samples from the same population of examinees we were in a situation where the use of the Mantel-Haenszel procedure to assess evidence of multidimensionality, as suggested by Rosenbaum (1984), seemed appropriate. In this use of the Mantel-Haenszel procedure the data are arrayed into these 2×2×K tables.

| Score on item i | Score on Item j: 1 | 0 | |
|---|---|---|---|
| 1 | $A_k$ | $B_k$ | $n_{1k}$ |
| 0 | $C_k$ | $D_k$ | $n_{0k}$ |
| | $m_{1k}$ | $m_{0k}$ | $T_k$ |

$k = 0, 1, \ldots, J-2 = K-1$ where k is the number-right score on the sum of all items except items i and j. If the test is unidimensional in the sense of being described by a latent variable model that a) is unidimensional, b) is locally independent, and c) has monotone item characteristic curves, then the MH D-DIF values computed for these 2×2×K tables must all be negative* (Rosenbaum, 1984) except for sampling error. In Rosenbaum's method one computes all of the MH D-DIF values for all possible pairs of items in the set of J being considered and examines this distribution for large positive values. If there are none, except for sampling variability, the item set has no evidence of multi-dimensionality. The more large positive MH D-DIF values there are in the distribution, the stronger the evidence that the J items are not unidimensional, at least for that population of examinees.

Table 15 gives the number and average of the positive MH D-DIF values that were significantly different from zero at the .05 level for each experimental and control section. It suggestively supports the multidimensionality theory in

---

Table 15 about here

---

*Rosenbaum does not use the ETS difficulty scale for the log of the common-odds-ratio estimate, while we do. This may cause some confusion since his logarithm will have the opposite sign of ours.

TABLE 15. MH D-DIF VALUES FOR DIMENSIONALITY FOR EXPERIMENTAL AND CONTROL SECTIONS.

| | | Average of Significantly Positive (.05) MH D-DIF Values | Number of Significantly Positive MH D-DIF Values |
|---|---|---|---|
| PS Sections | | | |
| Experimental | | | |
| | PSMin | .73 | 3 |
| | PSMax | 1.03 | 2 |
| Controls | | | |
| | F4 PS1 | .33 | 9 |
| | PS2 | .48 | 15 |
| | F5 PS1 | .79 | 3 |
| | PS2 | .56 | 5 |
| | F6 PS1 | .57 | 3 |
| | PS2 | .49 | 3 |
| SC Sections | | | |
| Experimental | | | |
| | SCMin | .88 | 3 |
| | SCMax | 1.07 | 2 |
| Controls | | | |
| | F4 SC | .42 | 15 |
| | F5 SC | .60 | 6 |
| | F6 SC | .54 | 9 |

the sense that the largest average MH D-DIF values are for the two maximum impact sections. The average for the two minimum impact sections are also large, but not as big as those for PSMax and SCMax. Unfortunately, these results are confounded with sample sizes. From Table 4 we observe that the four experimental sections were given to samples of about 1400 while the preoperational sections were given to twice that many examinees, and the operational controls to many times that number. Hence, the results of Table 15 are also compatible with the hypothesis that there is a) little evidence of multidimensionality in any of the sections, and b) the reason the experimental sections have the largest positive MH D-DIF value is that their estimates are much poorer, i.e., have the largest variances.

## 8. CONCLUSIONS

We will summarize our conclusions from this study as responses to a series of questions.

**A. Can tests be made that manipulate the differences in score distributions between Black and White examinees and yet maintain adherence to content specifications?**

Yes. Our experimental tests are parallel in form and content to other corresponding TEST sections and meet the content specifications for the assembly of PS and SC sections. These sections are not unusual insofar as an expert review of them as tests is concerned. And yet, the amount by which the Black/White score differences were manipulated ranged from 20 to 59 percent of the control average (Table 5).

**B. Can statistical specifications for such tests be met?**

Not sure. Our experimental sections did meet the targeted average difficulty level. But because of the relationship between impact, delta, and biserial correlations (Figures 1-4), the spread in difficulty levels of the

items and the average biserial correlation were either at the low or the high end of the acceptable ranges depending on whether the section was made of low or high impact items. These two differences, in turn, appear to alter the reliability of the experimental sections -- especially those assembled from low impact items. It is possible that by paying more stringent attention to the distribution of delta values (rather than simply their average value) and to the values of the biserial correlation coefficients that the experimental sections could have exhibited less variability in their reliability. Because of the relationship between impact, delta, and biserial correlations such increased stringency may have been impossible with the size of the item pools used in this study. At the very least it would have lowered the degree of impact manipulation that we achieved. Thus, we would predict that in order to meet the standards of reliability typically met in the assembly of the PS and SC sections either many more items must be available in the item pools or the amount of impact manipulation will be reduced to levels lower than those achieved in this study.

C. How much impact reduction is really possible?

Black examinees typically score about 1.1 standard deviations below the White examinees on the PS sections of the TEST. This would be changed to .9 standard deviations if PSMin were used. Similarly, the Black/White difference for SC sections on the TEST is .75 standard deviations and this would be reduced to about .5 standard deviations if SCMin were used. These changes in impact are substantial and several features of our study insured that they would be. We now consider three of them.

The number of Black examinees on which our initial item impact values were computed ranged from 250 to 350. This insured that the computed impact values

that we used to classify items would be stable and not subject to large sampling fluctuations. However, studies in which smaller numbers of Black examinees are used to compute item impact values (i.e., studies based on pretest data) will probably not result in as large a reduction in impact as we obtained here. This is because items that appear to have small impact based on small samples will tend to have larger impact values when they are used operationally, due to regression effects.

Because of the conflict between reducing impact and the traditional goal of maintaining test reliability (discussed in B) it is probable that, if the current standards of test reliability are to be met, items with somewhat more impact than those used in PSMin and and SCMin will need to be used in order to meet test specifications. This will lower the degree of impact reduction to levels below those we achieved here due to the trends relating item impact and item biserial correlations -- i.e., Figures 3 and 4.

Finally, the size of the item pool is an important constraint. SCMin, assembled from the smallest item pool (in relative terms) has the lowest reliability of all the sections of Table 9. In order to increase the reliability of SCMin it would probably be necessary to increase the size of our SC item pool by a factor of, perhaps, two (as it was for PSMin) or more.

In view of all these factors we believe that the degree of impact reduction found in this study is about as large as possible without significantly altering the statistical properties of the resulting tests and given the size of real item pools.

D. Does impact manipulation change the relationship of the resulting test scores to other variables?

Possibly. It is possible that the changes in the correlations discussed in section 6.3 could have been prevented if we had exercised more control over the

statistical characteristics of the experimental sections than we did. This possibility should be investigated in future work. If there is one strong lesson from this research it is that by introducing a new statistical characteristic into the test assembly process we run the risk of seriously affecting the standard statistical properties of the resulting tests. These statistical properties are the result of many years of research and test development and are closely tied to the validity of these tests as predictive instruments.

E. Are the minimum and maximum impact sections assembled in this study biased?

Yes. Clearly no test developer would ever assemble maximum impact sections for operational use because they are biased in a very clear-cut way. Our DIF analyses reported in section 7.1 also give evidence that maximum impact sections are biased against Black examinees -- i.e., the maximum impact sections had larger negative MH D-DIF values than the controls. Our minimum impact sections were developed using procedures that are like those proposed by advocates of the "Golden Rule" procedure whose goal is to reduce the disparity between the scores of Black and White (and other) examinees by mandating new, impact minimizing, test assembly procedures. However, our DIF analyses in section 7.1 suggest that PSMin and SCMin are biased against White examinees in exactly the same sense that PSMax and SCMax were biased against Black examinees.

F. What are the costs involved in impact reduction?

To achieve the level of impact reduction obtained in this study and to maintain the statistical properties of the resulting tests to insure their continued value as predictive instruments we would have needed larger item pools than we had. The PS pool was six times the size of a PS section and the SC pool was three times the size of an SC section. Item pools this much larger then current ones are impractical for many tests and thus the level of impact reduc-

tion will not, in practice, ever be as substantial as it was in this study. The costs involve increasing the size of the item pool, increasing the stringency of the statistical specifications of the test, or decreasing the target level of impact reduction.

An alternative approach might be to try to recover the loss of reliability that impact reduction creates by lengthening the test. Using the usual formula that relate reliability to test length (i.e., Lord and Novick, 1968, page 119) it may be shown that in order to raise the PSMin reliability of .71 to the control average of .77 we would need to lengthen the PS section by 37%. To raise the SCMin reliability of .58 to the control average of .69 we would need to lengthen the SC section by 61%. These values show that any attempt to counter the effects of impact reduction on reliablity by simply lengthening the test may involve rather substantial changes in total testing time, or, if not handed carefully, lead to more speeded tests.

G. <u>Is all impact manipulation achieved in the same way?</u>

No. The striking result of Table 6 and Figure 6 is that the manipulation of impact for PS is different from that of SC, in our data at least. For PS the scores of White examinees were unchanged and we altered the scores of the Black examinees. For the SC sections it was a trade-off, White scores went down and Black scores went up on SCMin while the reverse held for SCMax. These observations suggest that the impact reduction techniques used to assemble PSMin may have eliminated some items that disadvantaged Black examinees relative to White examinees. On the other hand, for SCMin these techniques may have replaced the eliminated items with those that disadvantage White examinees relative to Black examinees. These are intriguing results that need further clarification and evaluation on new data.

H. Are impact reduction techniques desirable additions to the test development process?

The results of this study shed some light on this complex question, but do not, of course, address all of the relevant issues. For example, we studied only discrete items (i.e., those that can all stand alone and do not, for example, depend on a common reading passage.) and only two types of items, PS and SC. In addition, our sample sizes were too small to allow us a thorough study of differential item functioning, omitting, and dimensionality, all of which would help us evaluate the quality of the minimum impact tests relative to the control tests.

Nonetheless, there are some fairly clear results of this study that ought to become part of the discussion of the use of impact reduction techniques in test development. For the item types and populations included in this study we may conclude the following.

First, such techniques can reduce impact if the initial impact values used to classify items are based on large enough samples of examinees.

Second, the resulting tests can be made to look parallel in form and content to conventionally constructed tests and meet their content specifications if the item pools are sufficiently large.

Third, the average difficulty level of the resulting tests can be maintained without changing current test development procedures for adhering to average difficulty specifications. However, the distribution of item difficulties will change if impact reduction techniques are used. There will be relatively more lower and higher difficulty items in the resulting tests and fewer middle difficulty items. This may be a controllable phenomenon, but will often require even larger item pools than used in this study, and will often

require test specifications that determine the distribution of item difficulty measures, rather than simply their average values. For some tests this effect is not too important. For tests with cut-points it is often desirable to concentrate the item difficulties near the cut-point and for such tests care will need to be exercised to insure that impact reduction techniques do not degrade the quality of the measurement of the test near the cut-point.

Fourth, tests constructed using impact reduction techniques will be less reliable than tests constructed conventionally. This also may be a controllable phenomenon but because of the positive relation between impact and item biserial correlations it is likely that there will always be a trade-off between maintaining test reliability and the amount of impact that can be reduced.

Fifth, the costs associated with impact manipulation techniques involve the need to increase current item pools by substantial factors, possibly as large as 3 to 6 times their current sizes, and to include a substantial pretesting activity in the test development process. For many testing programs these are impractical considerations. Furthermore, if only small increases in the size of item pools are possible the resulting impact reduction probably will be much smaller than what we achieved here.

Sixth, there is no clear sense in which tests in which impact has been reduced are better than those constructed conventionally. Certainly, our minimum impact sections were less reliable, for both Black and White examinees, than our control tests. Certainly, our minimum impact sections were not readily identified as improvements over our maximum impact and control sections by our expert reviewers. Our minimum impact sections did not have higher correlations with UGPA than the control sections. Finally, our DIF analyses in section 7.1 suggest that both SCmin and PSmin slightly favor Black examinees compared to

comparable scoring White examinees rather than being neutral in this regard.

As a final point we wish to reiterate the fact that the experimental test sections used in this study were non-operational and, even though real test takers responded to them, the performance of these examinees on the experimental sections did not contribute to their reported scores. In addition, the concern that the experience of taking a maximum impact section might discourage Black examinees and affect their performance on the rest of the test is not supported by the data given in Table 13. If Black examinees were discouraged by the questions in the Max sections, we would expect to see more omits in those sections than in the controls. Instead, for PSMax the omitting for Black examinees is less than it is for the controls while for SCMax it is about the same as for the controls.

## REFERENCES

Flaugher, R. and Schrader, W. (1978) "Eliminating differentially difficult items as an approach to test bias." Princeton, NJ: Educational Testing Service, ETS Research Bulletin RB-78-4.

Green, D. R. (1972) Racial and ethnic bias in test construction. Monterey, CA: CTB/McGraw-Hill.

Gulliksen, H. (1945) "The Relation of item difficulty and inter-item correlation to test variance and reliability." Psychometrika, 10, 79-91.

Holland, P. W. (1985) "On the study of Differential Item Performance without IRT." Proceedings of the Military Testing Association, October 1985, volume I, 282-287.

Holland, P. W. and Thayer, D. T. (1986) "Differential Item Performance and the Mantel-Haenszel Procedure." Presented at the American Educational Research Association Annual Meeting, April.

Ironson, G. H. and Craig, R. (1982) Item bias techniques when amount of bias is varied and score differences between groups are present. (Final report NIE-G-81-0045). Tampa, FL: University of South Florida.

Kok, F., Mellenbergh, G. and van der Flier, H. (1985) "Detecting experimentally induced item bias using the iterative logit method." Journal of Educational Measurement, 22, 295-303.

Linn, R. L. and Harnisch, D. L. (1981) "Interactions between item content and group membership on achievement test items". Journal of Educational Measurement, 18, 109-117.

Lord, F. and Novick, M. (1968) Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

McCallister, P. (1987) "Overview of state legislative initiatives concerning testing." Paper presented at the American Educational Research Association annual meeting, Washington, D.C., April, 1987.

Rosenbaum, P. R. (1984) "Testing the conditional independence and monotonicity assumptions of item response theory." Psychometrika, 49, 425-435.

Subkoviak, M., Mack, J., Ironson, G., and Craig, R. (1984) "Empirical comparison of selected item bias detection procedures with bias manipulation." Journal of Educational Measurement, 21, 49-58.

Suppes, P. and Zanotti, M. (1981) "When are probabilistic explanations possible?" Synthese, 48, 191-199.

# APPENDIX A

## The Influence of Variation in Item Characteristics on the Variance of Test Scores

## Appendix A

This analysis considers the variance of a number-right test score on a J item test. Because of the high correlation between number-right and formula scores, the results also holds for formula scores.

Let $X = \sum_{j=1}^{J} X_j$ denote the number right test score on a test of J items, where $X_j = 1$ for correct response to item j and $X_j = 0$ otherwise. Suppose a latent trait $\theta$ (which may be multivariate) underlies X, then let the item characteristic curve (ICC) of $X_j$ be

$$P_j(\theta) = P(X_j=1|\theta),$$

and let the average ICC be

$$\bar{P}(\theta) = \frac{1}{J} \sum_{j=1}^{J} P_j(\theta).$$

Theorem: If the $X_j$ are conditionally independent given $\theta$ then

$$\mathrm{Var}(X) = E\{J\,\bar{P}(\theta)\,\bar{Q}(\theta)\} + \mathrm{Var}\{J\,\bar{P}(\theta)\} - E\{\sum_{j=1}^{J}(P_j(\theta) - \bar{P}(\theta))^2\}$$

where $\bar{Q}(\theta)$ $1 - \bar{P}(\theta)$.

Proof:
$$\mathrm{Var}(X) = E\{\mathrm{Var}(X|\theta)\} + \mathrm{Var}\{E(X|\theta)\}.$$

But
$$\mathrm{Var}(X|\theta) = \mathrm{Var}(\sum_j X_j|\theta)$$

$$= \sum_j \mathrm{Var}(X_j|\theta)$$

$$= \sum_j P_j(\theta)\,Q_j(\theta)$$

where $Q_j(\theta) = 1 - P_j(\theta)$.

But
$$\sum_j P_j\,Q_j = \sum_j (P_j - \bar{P} + \bar{P})(Q_j - \bar{Q} + \bar{Q}) = \sum_j (P_j - \bar{P})(Q_j - \bar{Q}) + J\,\bar{P}\,\bar{Q}.$$

Hence
$$\sum_j P_j(\theta)\,Q_j(\theta) = J\,\bar{P}(\theta)\,\bar{Q}(\theta) - \sum_j (P_j(\theta) - P(\theta))^2.$$

Also $$E(X|\theta) = \sum_j P_j(\theta) = J\,\bar{P}(\theta).$$

Combining these results proves the Theorem □ .

Now consider the case where the ICCs are all equal. In that case the negative term in the theorem will be zero. Hence, the variance of X can only get smaller if the ICCs are unequal. This holds for all sets of ICCs with the same average value, $\bar{P}(\theta)$.

This is a precise result that shows why "peaking" a test will increase the variance of examinees scores while spreading out the ICCs, i.e., making some items easy and others hard but maintaining a fixed average difficulty level, will decrease the variance in examinee scores.

It is curious to observe that the assumption of conditional independence made in the theorem is always satisfied in the sense that for any random vector $X = (X_1, \ldots, X_J)$ there exists a one-dimensional latent variable $\theta$ for which the $X_j$ are independent given $\theta$ (Suppes and Zanotti, 1981). The ICCs are not necessarily monotone increasing but monotonicity of the ICCs plays no role in the theorem. When the ICCs are identical, the vector X has an exchangeable distribution -- but the converse is not true for finite J.